# The School Board is Nothing Like CNN: Challenges of Running Broadcast News Speech Components on Meeting Data

**David D. Palmer and Marc Reichman**
Virage Advanced Technology Group
300 Unicorn Park
Woburn, MA 01801
{dpalmer,mreichman}@virage.com

## Abstract

This paper describes our preliminary experiments in running an unadapted real-time broadcast news speech recognition system on a wide range of meeting data. We find the system performance in the output transcription to degrade from a range of 12-50% WER on English broadcast news sources to 70-120% WER on English meetings. We discuss several factors that contribute to this degradation.

## 1 Introduction

Over the past two decades, automatic speech recognition (ASR), or speech-to-text (STT), systems have been developed for and applied to a range of audio domains, with varying degrees of success. One of the most successful domains to which ASR systems have been applied has been broadcast news. There is growing interest in applying ASR systems to meeting data, including recent meeting corpus development (Janin *et al.* 2003) and evaluations of meeting transcription performance (Garofolo *et al.* 2004). The results presented in this paper represent both formal ASR evaluation results from this most recent NIST meeting evaluation and informal performance numbers from our internal experiments

Our interest in applying ASR systems to the meeting domain stems from our extensive work integrating real-time automatic speech and language processing systems into a large-scale news video indexing and retrieval system. As a preliminary step toward large-scale meeting processing, our goal in this work was to establish a performance baseline. Without modifying our existing real-time broadcast news ASR system, how would it perform on meeting data? From this baseline result, we would have a good idea of how much, and how, we might improve performance.

The Virage Enhanced Video Text and Audio Processing (eViTAP) system (Palmer *et al.* 2004) currently combines state-of-the-art automatic speech recognition and machine translation components with cross-lingual information retrieval in order to enable searching of multilingual video news sources by a monolingual speaker. In addition to full search capabilities, the system also enables real-time alerting, such that a user can be notified as soon as a word, phrase, or topic of interest appears in an English or Arabic news broadcast. Figure 1 shows the EViTAP cross-lingual search and alerting interface, with real data from the system, including both meeting data and broadcast news data. The list of relevant video clips matching an alerting profile or a user search is shown on the left, with broadcast source and time, most-frequent named entities, and a relevancy score. The center of the screen contains the video playback window, with clip navigation and keyframe storyboard. The right of the interface contains the transcripts from the ASR engine; video playback is synchronized with the transcripts such that words are highlighted as they are spoken in the video.

## 2 Meeting Data

As part of the ARDA Video Analysis and Content Extraction (VACE) program, we have been working with a set of meetings broadcast on the Reading, Massachusetts public access cable TV channel. This data set consists of a variety of Reading school board, town hall, and town council meetings that are open to the public. The meetings are typically recorded with a single camera and with multiple table-top and podium microphones combined into a single audio channel. The audio channel contains constant background noise, including paper shuffling, coughs, and side conversations. Topics of the meetings typically focus narrowly on school or town business, which is very different from the language data on which the broadcast news system was trained.

The data for the NIST meeting evaluation consisted of meetings from four sources, provided to NIST for use in the evaluation. The conversation topics range widely from music and politics to discussing the meetings themselves. There were three evaluation conditions in the formal NIST study, based on three different microphone recording types: Individual Headset Microphones (IHM), single distant microphones (SDM) and multiple distant microphones (MDM). Each of these conditions is different from the Reading Meetings we have been processing, which enabled us to analyze a broader set of meeting conditions.

## 3 ASR System Description

The ASR system we used for our experiments was a real-time broadcast news system developed by Softsound[1] that extends earlier BN speech work carried out at Cambridge University (e.g., Renals *et al.* 2000). The system has a vocabulary size of about 64,000 words, with pronunciations from the CMUDict (Weide 1998) or automatically generated from components of known words. The acoustic model was trained from about 24 hours of broadcast news data using HTK (Young, *et al.* 2000). Monophones were trained first, followed by top-down decision tree clustering, resulting in 3000 states and 16 Gaussians per mixture. A trigram language model was trained from data acquired through LDC/ELRA or from Internet spidering. The system uses a time-first decoder originally described in (Robinson and Christie 1998). The ASR system operates in 1X real time on all data.

## 4 Results

Speech recognition performance is measured in terms of a Word Error Rate (WER), which is the ratio of the total word substitutions, deletions, and insertions in the ASR system output to the total number of reference words in the transcript. The real-time Virage/Softsound ASR system typically produces a WER ranging from 12-50% on broadcast news data, with the best performance coming on “clean” anchor segments and the worst performance coming on “reporter in the field” segments.

As expected, the performance of the broadcast news system degrades significantly when applied to meeting data. Table 1 shows the system performance for each speaker in the Reading School Board data. The word error rate varies by speaker in the range 70-90%, with the overall performance for all speakers at 77.1%.

| Speaker | WER for BN system |
|---|---|
| Male 1 | 85.9% |
| Male 2 | 78.6% |
| Male 3 | 81.0% |
| Male 4 | 91.7% |
| Male 5 | 69.3% |
| Male 6 | 84.1% |
| Female 1 | 92.0% |
| All speakers | 77.1% |

Table 1: Performance of broadcast news ASR system for speakers in a school board meeting.

The performance of this system on the NIST meeting evaluation data for the single directional microphone (SDM) condition was roughly comparable, at 73.8%. This was consistent with our expectation, since the SDM condition was the most similar to the microphone environment at the school board meeting (a single audio recording representing input from multiple speakers or microphones).

The system performance on the individual head microphone (IHM) condition was significantly worse, at 116.0%. The IHM was the most controlled acoustic environment, and we expected the performance on this condition to be the best of all the data. Upon examining the results, we determined that there were two major factors in this unexpectedly-large degradation. The first was poor detection of long periods of silence in the data. The ASR system and audio classification system were trained on broadcast news data, in which silences are rarely longer than a second or two. In the meeting data, for certain microphones, silences were several minutes long. This

[1] A special thanks to Tony Robinson for making this system available for our experiments.

resulted in poor silence detection and misclassification of silences as speech (and thus large word insertion rates). The second factor in the performance degradation was the meeting ASR evaluation paradigm. For multiple IHM scenarios, the reference transcript for the data specified which single microphone was active at a given time, and all other microphones were assumed to be silent. However, our ASR system picked up background speakers on multiple microphones and (arguably correctly) transcribed these words, which were all treated as insertion errors in the evaluation. This effect can be seen clearly in Table 2, which shows the performance of the broadcast news system on the three meeting corpora plus the broadcast news baseline. The substitution and deletion error rate is comparable for the IHM and SDM conditions, but the insertion rate is extremely high for the IHM condition. We believe that this factor accounted for the majority of the degradation of the WER from 73 on SDM to 116 on IHM.

| Condition | S | D | I | WER |
|---|---|---|---|---|
| IHM | 31.1 | 34.5 | 50.4 | 116.0 |
| SDM | 30.4 | 41.5 | 1.9 | 73.8 |
| School Board | | | | 77.1 |
| Broadcast News | | | | 12-50 |

Table 2: Substitution, deletion, insertion, and word error rate performance of broadcast news ASR system for NIST meeting data, school board meeting, and broadcast news data.

## 5 Conclusions

Our preliminary experiments in applying a broadcast news ASR system to meeting data resulted in performance degradation from 12-50% WER on news data to 70-116% on meeting data. Not surprisingly, there are significant differences between broadcast news and meeting data. There is also a range of performance within each data type. However, across the range of meetings we processed, we observe a somewhat consistent degradation of ASR performance. It remains to be seen if this performance holds for additional types of meeting data, and it is likely to degrade further for meetings with lower-quality microphones or microphones that are not centrally located to all speakers.

In addition to our quantitative analysis of the ASR performance on meeting data, our experiments have resulted in the full integration of meeting video data, with full speaker ID and ASR transcripts, into our existing news processing system, as shown in Figure 1. This combination enables cross-lingual search, retrieval, and video playback by a system user across news and meeting domains.

## References

J. Garofolo, J. Fiscus, C. Laprun , “The RT-04 Spring Meeting Recognition Evaluation,” *Proceedings of ICASSP 2004 Meeting Recognition Workshop*, Montreal, 2004.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters , “The ICSI Meeting Corpus, ” *Proceedings of ICASSP-2003*, Hong Kong, April 2003.

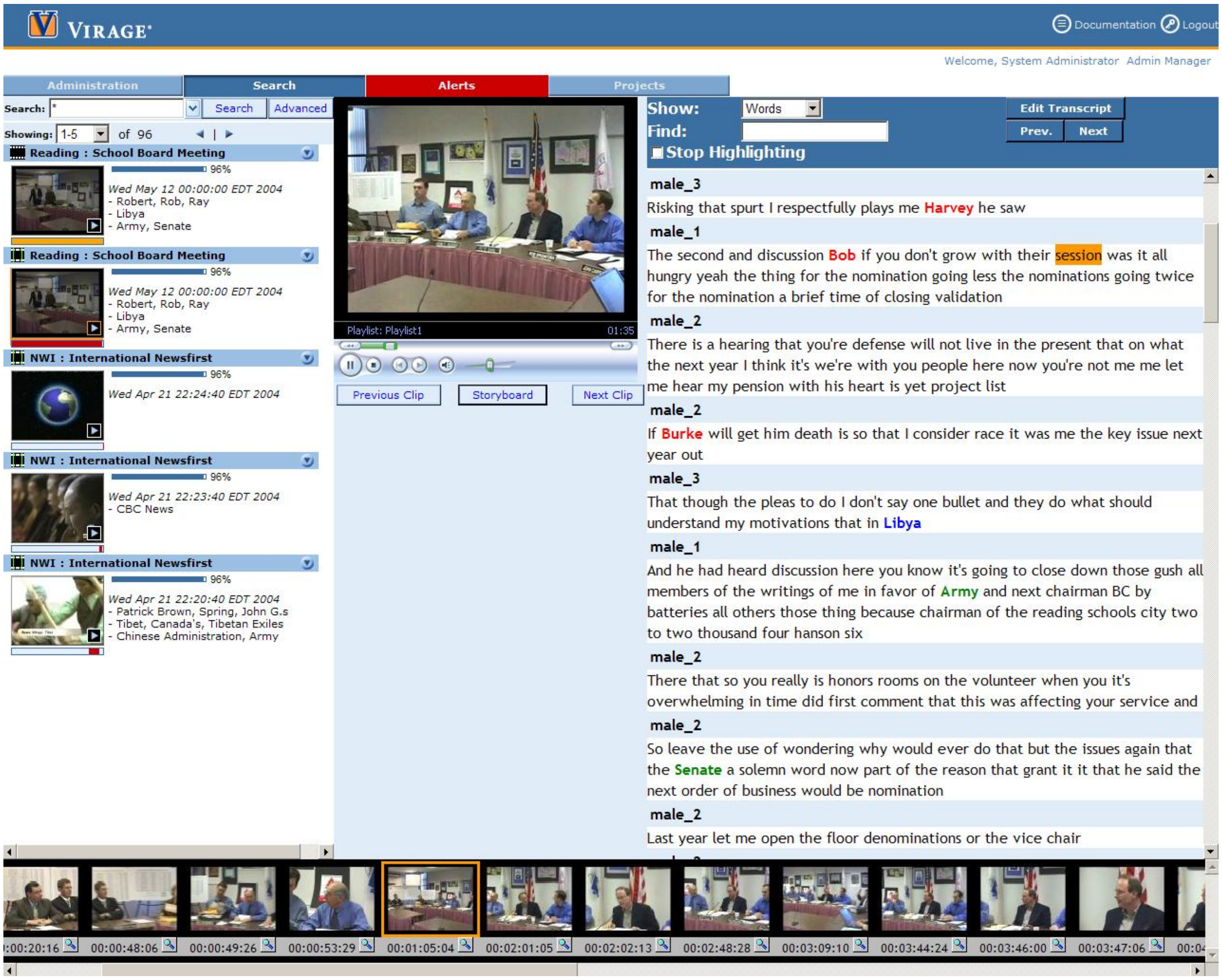
D. Palmer, P. Bray, M. Reichman, K. Rhodes, N. White, A. Merlino, F. Kubala, “Multilingual Video and Audio News Alerting,” In *HLT/NAACL 2004: Companion Volume*, 2004.

S. Renals, D. Abberley, D. Kirby, and T. Robinson, “Indexing and Retrieval of Broadcast News,” Speech Communication, Vol. 32, pp. 5-20, 2000.

T. Robinson and J. Christie, “Time-first search for large vocabulary speech recognition,” In *Proc. IEEE ICASSP*, pages 829-832, 1998.

R. Weide. *Carnegie Mellon Pronouncing Dictionary*. http://www.cs.cmu.edu, 1998.

S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and P. Woodland. "The HTK Book Version 3.0." Cambridge, England, Cambridge University, 2000.



Reading : School Board Meeting
96%
Wed May 12 00:00:00 EDT 2004
- Robert, Rob, Ray
- Libya
- Army, Senate

Reading : School Board Meeting
96%
Wed May 12 00:00:00 EDT 2004
- Robert, Rob, Ray
- Libya
- Army, Senate

NWI : International Newsfirst
96%
Wed Apr 21 22:24:40 EDT 2004

NWI : International Newsfirst
96%
Wed Apr 21 22:23:40 EDT 2004
- CBC News

NWI : International Newsfirst
96%
Wed Apr 21 22:20:40 EDT 2004
- Patrick Brown, Spring, John G.s
- Tibet, Canada's, Tibetan Exiles
- Chinese Administration, Army

**male_3**
Risking that spurt I respectfully plays me **Harvey** he saw

**male_1**
The second and discussion **Bob** if you don't grow with their session was it all hungry yeah the thing for the nomination going less the nominations going twice for the nomination a brief time of closing validation

**male_2**
There is a hearing that you're defense will not live in the present that on what the next year I think it's we're with you people here now you're not me me let me hear my pension with his heart is yet project list

**male_2**
If **Burke** will get him death is so that I consider race it was me the key issue next year out

**male_3**
That though the pleas to do I don't say one bullet and they do what should understand my motivations that in **Libya**

**male_1**
And he had heard discussion here you know it's going to close down those gush all members of the writings of me in favor of **Army** and next chairman BC by batteries all others those thing because chairman of the reading schools city two to two thousand four hanson six

**male_2**
There that so you really is honors rooms on the volunteer when you it's overwhelming in time did first comment that this was affecting your service and

**male_2**
So leave the use of wondering why would ever do that but the issues again that the **Senate** a solemn word now part of the reason that grant it it that he said the next order of business would be nomination

**male_2**
Last year let me open the floor denominations or the vice chair

Figure 1: Virage Enhanced Video Text and Audio Processing (eViTAP) interface, showing school board and broadcast news data, with speaker ID and speech recognition output synchronized with video playback.